UNCLASSIFIED

Defense Technical Information Center
Compilation Part Notice

ADP010393

TITLE: Language Adaptive LVCSR Through Polyphone
Decision Tree Specialization

DISTRIBUTION: Approved for public release, distribution unlimited

This paper is part of the following report:

TITLE: Multi-Lingual Interoperability in Speech
Technology [l'Interoperabilite multilinguistique
dans la technologie de la parole]

To order the complete compilation report, use: ADA387529

The component part is provided here to allow users access to individually authored sections of proceedings, annals, symposia, ect. However, the component should be considered within the context of the overall compilation report and not as a stand-alone technical report.

The following component part numbers comprise the compilation report:
ADP010378 thru ADP010397

UNCLASSIFIED

Language adaptive LVCSR through Polyphone Decision Tree Specialization

T. Schultz[1] *and A. Waibel*[2]
{tanja@ira.uka.de}

[1] Interactive Systems Laboratories
University of Karlsruhe
Aussfarengarten 5a
76131 Karlsruhe, Germany

[2] Universität Karlsruhe, ILKD
Am Fasenengarten 5
D-76128 Karlsruhe, Germany

ABSTRACT

With the distribution of speech technology products all over the world, the fast and efficient portability to new target languages becomes a practical concern. In this paper we explore the relative effectiveness of porting multilingual recognition systems to new target languages with very limited adaptation data. For this purpose we introduce a polyphone decision tree specialization method. Several recognition results are presented based on mono- and multilingual recognizers developed in the framework of the project GlobalPhone which investigates LVCSR systems in 15 languages.

Language	Abbr	Utts	Spks	Units	Hours
Ch-Mandarin	CH	8529	112	219K	26.7
Croatian	CR	3374	72	89K	12.0
English (WSJ)	EN	7137	83	129K	15.0
French (Bref)	FR	7143	74	123K	13.9
German	GE	9173	71	132K	16.7
Japanese	JA	9096	108	212K	22.9
Korean	KO	6335	80	301K	16.4
Spanish	SP	5419	82	138K	17.6
Turkish	TU	5466	79	87K	13.2
Total		68276	839	1554K	170.4

Table 1: GlobalPhone database used for experiments

1. Introduction

With the distribution of speech technology products all over the world, the fast and efficient portability to new target languages becomes a practical concern. So far one major time and costs limitation in developing LVCSR systems in new languages is the need of large training data. According to the amount of data used for porting acoustic models to a new target language we differentiate three aspects of research:

- ⋆ Cross-language transfer (no data)
- ⋆ Bootstrapping (much data)
- ⋆ Language adaptation (very limited data)

The term *cross-language transfer* refers to the technique where a system developed in one language (group) is applied to recognize another language without using any training data of the new language. We do not distinguish whether the transfer to the target language is done from one language or from a group of languages. Research focuses on the questions whether cross-language transfer from one language to another language of the same family performs better than across family borders [4], and second if the number of languages used for training the transfer models influences the performance on the target language [7], [13]. Results seems to indicate a relation between language similarity and cross-language performance [4], [3]. Furthermore it is clearly shown that multilingual transfer models outperform monolingual ones [3], [14].

The key idea in the *bootstrapping* approach is to initialize a recognizer in the target language by using already developed acoustic models from other language(s) as seed models. After the initialization step the resulting system is completely rebuild using large training data of the target language. This idea was first proposed by Zue and evaluated by [6] and [15] showing that crosslanguage seed models perform better than flat starts or random models. Recently the usefulness of multilingual phonemic inventories and multilingual phoneme models as seed models have been demonstrated by [9], [11].

The *language adaptation technique* lies between the two extremes in terms of available training data. In this approach an existing recognizer is adapted to the new target language with only very limited data. [15],[9], [10] focus on two issues: first the amount of data needed to get reasonable results, second the question of finding suitable acoustic models to start from. For the first question they found -coincident to our expectation- that the language adaption performance is strongly related to the amount of data used for adaptation. [15] proved that the number of different speakers used for training is more critical than the number of utterances. The question of suitable models to start from was investigated by [9] and [10] comparing the effectiveness of multilingual acoustic models. Again it could be shown that multilingual models outperform monolingual ones.

Previous systems which combined multilingual acoustic models have been limited to small tasks and context independent modeling. Since for the monolingual case the use of larger phonetic context windows has proven to increase the recognition performance significantly, such improvements extend naturally to the multilingual setting. The idea how to construct context dependent multilingual models was first proposed by [5] and [14]. For the language adaptation purpose we intend to exploit the context information learned from several lan-

guages. How this information can be incorporated into the language adaptive process is still an open issue. In this paper we present a new approach to adapt polyphone decision trees to the new target language.

2. Multiple Languages

For our experiments we developed monolingual LVCSR systems in nine languages which will be introduced in this section. For training and testing we are using our multilingual database GlobalPhone.

2.1. The GlobalPhone Database

GlobalPhone currently consists of the languages Arabic, Chinese (Mandarin and Shanghai dialects), Croatian, German, Japanese, Korean, Portuguese, Russian, Spanish, Swedish, Tamil, and Turkish. In each of these languages we collected about 15 hours of speech spoken by 100 native speakers per language. Every speaker read several articles from a national newspaper. The articles were chosen from the areas: national politics, international politics, and economy. The speech data was recorded at a sampling rate of 48kHz using a close-talking microphone connected to a DAT-recorder. After transferring the sound data from DAT to hard disc it was downsampled to 16kHz, 16-bit. The GlobalPhone corpus is fully transcribed, and during validation process special markers were added for spontaneous effects like false starts, and hesitations. Further details about the GlobalPhone project are given in [12].

Since English and French are already available in very similar frameworks we decided not to collect additional data in these well covered languages but add the two databases Wall Street Journal (WSJ0, distributed by LDC) for English and Bref (BREF-Polyglot sub-corpus, distributed by Elsnet) for French to our training data. The resulting database covers 9 of the 12 most widespread languages of the world.

Throughout the experiments 80% of the speakers were used for training the acoustic models, 10% were defined as a test set, and the remaining 10% were kept as further cross-validation set. See table 1 for an overview of the database used throughout the experiments.

2.2. Monolingual Baseline Recognizers

We developed equally designed monolingual LVCSR systems in nine languages using our Janus Recognition Toolkit (JRTk). For each language the resulting baseline recognizer consists of fully continuous 3-state HMM systems with 3000 polyphone models. Each HMM-state is modeled by one codebook which contains a mixture of 32 Gaussian distributions. The preprocessing is based on 13 Mel-scale cepstral coefficients with first and second order derivatives, power and zero crossing rate. After cepstral mean subtraction a linear discriminant analysis reduces the input to 32 dimensions.

Language	Word based			Phoneme based		
	ER	Vocab	PP	ER	Vocab	PP
Ch-Mandarin	14.5	45K	207	45.2	141	12.5
Croatian	20.0	15K	280	36.7	32	9.6
English	14.0	64K	150	46.4	46	9.2
French	18.0	30K	240	36.1	38	12.1
German	11.8	61K	200	44.5	43	9.0
Japanese	10.0	22K	230	33.8	33	7.9
Korean	31.0	64K	130	36.1	43	9.9
Spanish	20.0	15K	245	43.5	42	8.2
Turkish	16.9	15K	280	44.1	31	8.5

Table 2: Word and phoneme based error rates (ER), vocabulary size, and trigram perplexity (PP) for nine languages

In table 2 we arranged the error rates[1], vocabulary size and trigram perplexities for the monolingual recognizer. Since the engines are the same across the languages, differences in the recognition performance are due either to language specific inherent difficulties or to differences in quality and quantity of the used knowledge sources and data. In our opinion it is misleading to infer from the given word error rates to language difficulties. On the one hand the concept of a word does not hold for each language (Chinese, Japanese, and Korean). On the other hand the word error rates are strongly affected by available corpus data and resulting artifacts like different vocabulary sizes, OOV-rates, language model perplexities, and last but not least by the human language expertise, which in our case is not comparable in all languages.

A reliable measure of the acoustic difficulties of the nine languages is the phoneme based recognition rate using a phoneme recognizer without any (phoneme) language model constraints. The results in table 2 indicate significantly differences in acoustic confusability between languages, ranging from 33.8% to 46.4% phoneme error rate. English seems to be the most hardest task in acoustical sense whereas Japanese is the easiest.

3. Multilingual Systems

In this section we describe our approach to create a multilingual recognizer engine by combining context dependent acoustic models across languages.

3.1. Global Phonetic Inventory

We intend to share acoustic models of similar sounds across languages for the adaptation purpose. Those similarities can be either derived from international phonemic inventories documented in Sampa, Worldbet, and IPA or by data-driven methods as proposed for example by [1].

In our work we defined a *global phoneme set* based on the phonemic inventory of the monolingual systems. Sounds

[1] Mandarin is given in character based error rate, Japanese in hiragana based error rate, and Korean in syllable based error rate

which are represented by the same IPA symbol share one common phoneme category. In case of five languages we started with 171 language specific phonemes and pooled them together into 85 phoneme categories. In case of nine languages we pooled 339 language dependent phonemes into 140 phoneme categories. Thus the phone-set compression rate of 49% in the five-lingual case increases to 41% in the nine-lingual case.

3.2. Multilingual acoustic model Combination

Based on the above described phoneme categories we designed multilingual systems by combining the language dependent acoustic models of the languages Croatian, Japanese, Korean, Spanish, and Turkish in two different ways and compared their effectiveness for the language porting purpose.

In system *ML-mix* we share all models across these five languages without preserving any language information. We build context dependent models by applying a decision tree clustering procedure which uses a question set of linguistic motivated phonetic context questions. We train the models by sharing the data of the five languages. In the second system *ML-tag* the phoneme model sharing across languages is performed by attaching a language tag to each of the phoneme categories in order to preserve the information about the language. The above described clustering procedure is enhanced by introducing questions about the language and language groups to which a phoneme belongs. Therefore the decision if phonetic context information is more important than language information becomes data-driven (see [14] for details).

We explore the usefulness of the two different modeling approaches by running three experiments on 7 recognizers summarized in table 3:

1. Monolingual baseline test: all five monolingual recognizers are tested on the corresponding language
2. Multilingual test: *ML-mix* and *ML-tag* are applied to recognize one of the five languages involved in training the multilingual models
3. Porting test: the five monolingual systems as well as *ML-mix* and *ML-tag* are applied to recognize German utterances.

The results of the multilingual test show that *ML-tag* outperforms the mixed system *ML-mix* by 5.3% (3.1% - 8.7%) error rate. This indicates that preserving the language information achieves better results with respect to the ideal situation that sufficient training data is available to build a language specific system. This finding is coincident to other studies [5], [9]. The porting test prove that *ML-mix* outperforms *ML-tag* in both techniques. This is evident since sharing information across languages augments the language robustness of the transfer system (see [14] and [10] for details).

Language	Mono	ML-tag	ML-mix
Croatian	26.9	31.9	35.0
Japanese	13.0	15.0	20.0
Korean	47.3	49.0	
Spanish	27.6	32.4	37.0
Turkish	20.1	21.3	29.0
Technique	Porting to German		
Crosslanguage	49.5-65.0	50.0	41.5
Bootstrap	28.4-50.5	35.7	29.2

Table 3: Word error rates of *ML-mix* versus *ML-tag*

3.3. Dictionary Mapping

For all our experiments we presume that a pronunciation dictionary for the target language is given in an arbitrary phoneme set. Since we are interested in time and cost effective algorithms we created dictionaries which are not already available from scratch by grapheme-to-phoneme tools. However we post-edit the dictionaries by human experts who added pronunciation variants and treated special events like acronyms.

Nevertheless for recognizing the target language with the *ML-mix* or *ML-tag* system we need to define an appropriate mapping from our global phoneme set to the target phonemes. We investigate two approaches to find this mapping: In the first approach we apply an heuristic IPA-based mapping, meaning that a human experts defines for each target phoneme the corresponding counterpart according to our IPA phoneme categories. In the second approach we perform a data-driven mapping by calculating a phoneme confusion matrix, and picking the phoneme as a counterpart which leads to the highest confusion with the target phoneme. For this experiment we assume that an accurate phoneme recognizer in the target language is already given. We calculated phonetic alignments of 500 utterances spoken in the target language and did a frame-wise comparison with the viterbi decoded alignment of the same 500 utterances using a multilingual recognizer. Our experiments show that the IPA-based approach outperforms the data-driven approach by 27.1% vs 34.3% word error rate for the bootstrap technique and 66.7% vs 74.5% word error rate for the cross-language transfer technique (see [13] for details).

4. Polyphone Decision Tree Specialization

When creating the *ML-mix* system we uses a divisive clustering algorithm that builds context querying decision trees [8]. As selection measure for dividing a cluster into two subclusters we used the maximum entropy gain on the mixture weight distributions. This clustering approach gave significant improvements across different tasks and languages [8]. Figure 1 shows for 10 languages the number of different models we can get when using different context sizes. As can

be seen these numbers differs very much between the languages. These differences are due to the perplexity of the language, to the number of words in the training corpus, and to the length of the modeled word units. The latter is according to a contraint imposed by the decoder which limits the maximum context width to all phonemes within a word and up to one phoneme into the neighboring words. For example the extremely shortness of Korean units used in our recognizer results in zero polyphones of context larger than 2. While for Chinese, Japanese and Spanish the most frequent word length in the training data is 2 phonemes, it is 5 for Turkish and 6 for Russian. The most frequent numbers of phonemes in the dictionary various from 2 for Spanish to 9 for Turkish.

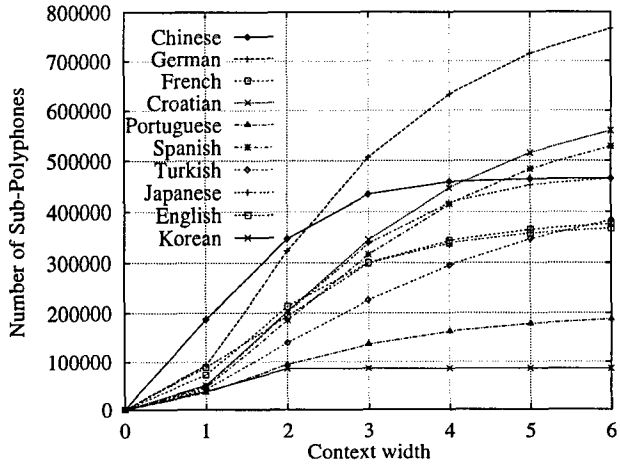

Figure 1: Different Sub-polyphones in training corpus

The concept of the IPA-based phoneme categories allows us to share context models across languages. To estimate the percentage of polyphone overlap between languages we define the non symmetric polyphone coverage measure as the number of polyphone occurrences in one language covered by polyphones in another language. In table 4 we give the triphone coverage for 10 languages. Here we distinguish between the coverage of polyphone types (upper row) and the coverage of polyphone occurrences (lower row), where the first one focus on the aspect whether common polyphones exists across languages, and the latter one focus on the aspect that frequent polyphones are more important to cover than rare ones. For example 33.6% of Japanese triphone occurrences are covered by German triphones, whereby 22.3% of the polyphone types are responsible for this coverage rate. On the other hand only 19.5% of German triphone occurrences are covered by Japanese polyphones. This effect is due to the Japanese phonotactic which only allows consonant vowel combinations.

From table 4 it is obvious that we should be aware of a large mismatch between represented polyphones in the multilingual

B/C	CH	DE	EN	FR	JA	KO	KR	PO	SP	TU
CH	100	0.1	0.3	0.1	0.1	0.0	0.1	0.1	0.1	0.2
		5.3	6.8	5.8	4.2	5.3	4.2	5.4	5.3	4.9
DE	0.1	100	5.5	19.8	9.3	7.2	18.6	13.6	12.9	12.9
	3.9		19.6	41.6	19.5	18.2	34.9	28.0	28.3	26.1
EN	0.6	5.4	100	6.5	1.8	3.4	1.5	0.9	1.3	3.8
	5.2	18.1		18.6	8.9	11.6	7.7	6.6	6.6	9.2
FR	0.1	29.0	9.7	100	10.2	11.2	25.8	18.4	17.4	23.1
	3.9	53.3	16.4		22.7	28.7	45.5	36.4	41.3	35.6
JA	0.2	22.3	4.5	16.8	100	9.8	16.0	11.0	13.6	25.9
	2.5	33.6	9.9	37.4		25.6	29.2	27.6	31.2	52.5
KO	0.1	10.3	4.9	10.9	5.8	100	10.2	8.0	9.3	9.1
	4.1	36.3	16.1	35.0	24.9		38.6	30.8	38.4	26.1
KR	0.2	39.0	3.2	37.0	14.0	15.0	100	31.0	34.3	31.5
	1.8	68.8	5.0	64.7	28.2	34.5		63.0	61.8	50.4
PO	0.4	30.2	2.0	28.0	10.2	12.5	32.9	100	33.5	19.8
	2.3	57.9	4.6	49.5	26.7	37.5	62.5		57.5	39.9
SP	0.2	25.4	2.7	23.5	11.2	12.9	32.2	29.7	100	17.5
	2.5	60.2	5.6	60.1	34.0	40.1	64.2	58.2		41.0
TU	0.8	29.6	8.9	36.3	24.8	14.6	34.4	20.4	20.3	100
	5.4	46.0	18.3	52.0	46.1	33.0	50.1	38.6	39.6	

Table 4: Triphone Coverage for 10 languages

decision tree and the observed polyphones in a new target language. We therefor specialize the already existing multilingual polyphone decision tree to the new language by continuing growing the decision tree. The limited amount of adaptation data is used to train separate mixture weight distribution for the resulting leaf nodes.

5. Language Adaptation to Portuguese

In the previous sections we report on the usefulness of multilingual acoustic model combination with respect to porting these acoustic models to the German language with the cross-language transfer and bootstrap technique. Now we investigate the benefit of these multilingual models in combination with the polyphone decision tree specialization (PDTS) for language adaption. We intend to adapt the different described multilingual systems to Portuguese. For adaptation we presume that a Portuguese dictionary as well as the recordings and transcriptions of 200 spoken utterances are given. Although [15] found that the number of speakers for adaptation is more critical than the number of utterances we decide to use 200 utterances spoken by only 7 different Portuguese speaker since at least in our dictation task it is more expensive to get single utterances of many different speakers than to get many utterances spoken by one speaker. The 200 utterances result in 25 minutes speech with 3370 spoken word units for adapting the acoustic models. The dictionary mapping was done according to our heuristic IPA-based mapping approach.

A subset of 96 uniformally selected utterances from 3 test speakers was used to carry out our experiments. The test dictionary has 7300 entries, the OOV-rate is set to 0.5% by including the most common words of the test set into the dictionary. A trigram language model with Kneser/Ney backoff

scheme was calculated on 10 million word text corpus from Agency France Press interpolated with the GlobalPhone data leading to a trigram perplexity of 297.

5.1. Polyphone Coverage

Before applying our polyphone decision tree specializing approach we want to examine how well the 49 Portuguese monophones and resulting polyphones are covered by the nine- and five-language pool. Therefore we calculated the coverage of Portuguese polyphones according to our IPA phoneme categories. This measure indicates how well a not specialized polyphone decision tree fits to the target language. The coverage is shown in figure 2 for context width 0 (monophones) and 1 (triphones). The calculation of plotted coverage proceeds as follows: first we select that language among all pool languages which achieves the highest coverage for Portuguese. We then remove this language from the pool and calculate the coverage between Portuguese and each language pair resulting from the combination of removed language plus remaining pool language. The procedure is repeated for triples and so forth. Thus in each step we find the language which maximally complements the polyphone set.

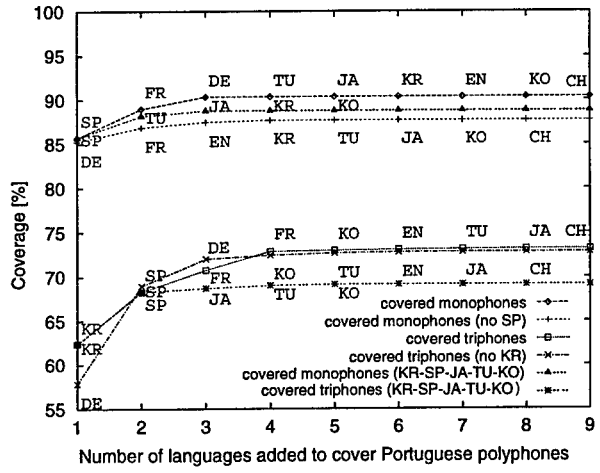

Figure 2: Portuguese polyphone coverage by nine languages

From the figure 2 we observed that as expected the coverage dramatically decrease for larger context (for quintphones a maximal coverage of 46% could be attained). After incorporating three languages the coverage of Portuguese monophones can not increased any further, limited to 91% with the nine language pool and dropping to 85% when the most important language for monophone coverage (SP) is removed from the language pool. The contribution of the Spanish phoneme set to the monophone coverage can not be compensate by other languages remaining in the pool. Second we found that when increasing the context width to 1 the coverage saturate after four languages. When increasing to context width to 2 we observed that at least five languages contribute to the quintphone coverage rate. Therefor we infer that increasing the context width requires more languages. For the context width 1 the main contribution comes from the Croatian language. Removing this language from the pool is nearly completely compensate by German and Spanish triphones. This indicate that Croatian, German, and Spanish polyphones covers a similar portion of the Portuguese triphones set. Whereas the curve (KR-SP-JA-TU-KO) indicates that the French language contribute unique polyphones which can not be recruited from other languages. In this case the lacking phonemes belong to the categories of nasal vowels. We conclude from this observation that when designing a language pool for adaptation purposes it is more critical to find a complement set of languages than to cover a large number of languages. Our method of calculating the polyphone coverage across languages can help to find such a complementary language set. From analyzing the polyphone coverage we draw the conclusion that using a polyphone tree even based on several languages can not be applied successful to Portuguese without adapting to the new contexts.

System	Data	Labels	Technique	Ptree
	Cross-language transfer			
S1	0	-	-	ML
S2	0	-	-	CI
	Language adaptation			
S4	100	initial	MLAdapt	CI
S5	100	initial	Viterbi	ML
S6	100	initial	MLAdapt	ML
S7	100	good	MLAdapt	ML
S8	200	good	MLAdapt	ML
S9	200	good	PDTS	ML-PO
	Bootstrap			
S3	100	initial	Rebuild	PO
S10	6600	good	Rebuild	PO

Table 5: Description of systems adapted to Portuguese

5.2. Results

Table 5 describes the systems used for our adaptation experiments, their performance on Portuguese is compared in figure 3. The column **Data** in table 5 refers to the number of recordings used as adaptation data. Applying no data results in a cross-language transfer approach as performed in the systems S1 and S2. Whereas the training based on 6600 utterances (S10) represents the bootstrap technique. For the systems S3 to S9 we used very limited data of 100 and 200 utterances.

Labels explains whether the phonetic transcription of the recordings are created based on the multilingual recognition engine *ML-mix* (Labels = initial) or based on good phonetic alignments which we presume to be already given (Labels = good). The latter was used to accelerate our adaptation process. In future work we will examine if we can get close to this label quality by iterating our adaptation approach.

The term **Technique** is related to the training approach applied to the systems. Viterbi refers to one iteration of viterbi training along the given labels. MLAdapt means Maximum Likelihood Adaptation technique, Rebuild refers to the iterative procedure of writing labels, viterbi training, model clustering, training, and writing improved labels. PDTS is the described Polyphone Decision Tree Specialization.

The **Ptree** item describes the origin of the polyphone decision trees. CI refers to context independent modeling, meaning that no polyphone tree is used, ML is the 3000 polyphone tree of system *ML-mix* and PO is a polyphone tree build exclusively on Portuguese polyphones. ML-PO refers to the regrown *ML-mix* polyphone tree applying PDTS.

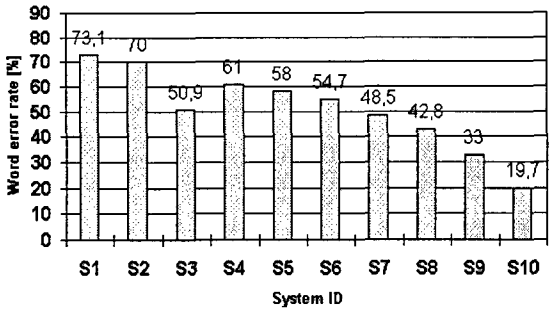

Figure 3: Language adaptation to Portuguese

As expected the recognition of Portuguese speech by running the five-lingual recognizer *ML-mix* without any training data results in extremely high word error rates of 73.1% for the context dependent system (S1) and slightly better error rates of 70% for the context independent system (S2). Therefor the initial labels are written with the multilingual context independent system S2. Using 100 of these initial labels for adapting the context independent multilingual system (S4) and the context dependent system by MLA (S6) or viterbi training (S5) shows a significant gain. In S3 the initial labels are used to completely rebuild a Portuguese system after bootstrapping from multilingual seed models. The comparison of S6 and S3 indicate that the adaptation of non matching polyphone trees is outperformed by the bootstrap technique (S3) even if data are very limited. Nevertheless the word error rate of the winning system S3 achieving 50.9% is still unsatisfying.

We obtain the next performance boost from using improved labels (S7) and double amount of adaptation data (S8). Finally we applied our PDTS approach (S9) which leads to significant improvements achieving 33% word error rate. This performance compares to 19.7% word error rate (S10) resulting from bootstrapping and rebuilding a Portuguese LVCSR system using 16 hours of speech spoken by 78 speakers. To summarize we get the highest performance gain in language adaptation from the PDTS technique, enlarging adaptation data, and improved labels, in this order.

6. Conclusion

In our language adaptive approach we explore the relative effectiveness of multilingual context dependent acoustic models in combination with a polyphone decision tree specialization (PDTS). We examine the profit when porting a multilingual engine to new target languages with very limited training data. The results are very promising achieving 33% word error rate for an Portuguese LVSCR system when using only 200 spoken utterances for adaptation.

7. Acknowledgment

The authors gratefully acknowledge all members of the GlobalPhone team for their great enthusiasm. We also wish to thank the members of the Interactive Systems Laboratories, especially Roald Wolff for his active support, great encouragement and contribution to this research.

References

1. O. Andersen et al.: *Data-Driven identification of Poly- and Mono-phonemes for four European Languages* in: Proc. Eurospeech, pp. 759-762, Berlin 1993.
2. J. Barnett et al.: *Multilingual Speech Recognition at Dragon Systems* in Proc. ICSLP, pp. 2191-2194, Philadelphia 1996.
3. U. Bub et al.: *In-Service Adaptation of Multilingual Hidden-Markov-Models*, Proc. ICASSP, pp. 1451-1454, Munich 1997.
4. A. Constantinescu et al.: *On Cross-Language Experiments and Data-Driven Units for ALISP* in: Proc. ASRU, pp. 606-613, St. Barbara, CA 1997.
5. P. Cohen et al.: *Towards a Universal Speech Recognizer for Multiple Languages* in: Proc. ASRU, pp. 591-598, St. Barbara CA, 1997.
6. J. Glass et al.: *Multi-lingual Spoken Language Understanding in the MIT Voyager System* in: Speech Communication (17), pp. 1-18, 1995.
7. S. Gokcen et al.: *A Multilingual Phoneme and Model Set: Towards a universal base for Automatic Speech Recognition* in: Proc. ASRU, pp. 599-603, St. Barbara, CA 1997.
8. M. Finke et al.: *Wide Context Acoustic Modeling in Read vs. Spontaneous Speech* in: Proc. ICASSP, Munich 1997.
9. J. Köhler: *Language Adaptation of Multilingual Phone Models For Vocabulary Independent Speech Recognition Tasks*, Proc. ICASSP, pp. 417-420, Seattle, 1998.
10. T. Schultz et al.: *Language independent and language adaptive LVCSR* in: Proc. ICSLP, pp. 1819-1822, Sydney 1998.
11. T. Schultz et al.: *Fast Bootstrapping of LVCSR Systems with Multilingual Phoneme Sets* in: Proc. Eurospeech, pp. 371-374, Rhodes 1997.
12. T. Schultz et al.: *The GlobalPhone Project: Multilingual LVCSR with Janus-3* in: Proc. SQEL, pp. 20-27, Plzeň 1997.
13. T. Schultz et al.: *Adaptation of Pronunciation Dictionaries for Recognition of Unseen Languages* in: Proc. Specom, pp. 207-210, St. Petersburg, Russia 1998.
14. T. Schultz et al.: *Multilingual and Crosslingual Speech Recognition* in: Proc. DARPA Workshop on Broadcast News Transcription and Understanding, Lansdowne, VA 1998.
15. B. Wheatley et al.: *An Evaluation of Cross-language Adaptation For Rapid HMM Development in a new language* in: Proc. ICASSP, pp. 237-240, Adelaide 1994.